Evaluating the Master Address File—Auxiliary Reference file (MAF-ARF) as a Potential Respondent Retention Source

by

Michaela Dillon
U.S. Census Bureau

CARRA Working Paper 18-12 **September, 2021**

This paper was written in 2018 under the former Center for Administrative Records Research and Applications program, but not published until 2021. CARRA has since been combined with the Center for Economic Studies, and the CARRA working paper series has been discontinued.

To obtain information about the series, see www.census.gov/ces or contact Christopher Goetz, Editor, Discussion Papers, U.S. Census Bureau, Center for Economic Studies 5K038E, 4600 Silver Hill Road, Washington, DC 20233, CES.Working.Papers@census.gov.

Abstract

Using administrative records in survey operations can potentially improve data accuracy and survey operations. In this study, we link administrative data on residential location from the Master Address File-Auxiliary Reference File (MAF-ARF) dataset to the National Survey of College Graduates (NSCG) to understand the alignment of this administrative records (AR) information with respondent collected data. The agreement rate for both sources reporting the same address varies from around 85 percent in 2011 to less than 60 percent in 2015. Between the 2010 and 2013 surveys, the MAF-ARF predicts just over half (54.54 percent) of non-returning respondents did not move from their initially reported address, suggesting that respondent refusals to continue participation as opposed to inability to locate them may be more important for understanding unit non-response.

Keyword*:* administrative records, college graduates, sampling frames

JEL Classification: C83, E2, C8

[*] This paper is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed, DRB approval number CBDRB-FY18-431.

EXECUTIVE SUMMARY

Using administrative records in survey operations can potentially improve data accuracy and survey operations. In this study, we link administrative data on residential location from the Master Address File-Auxiliary Reference File (MAF-ARF) dataset to the National Survey of College Graduates (NSCG) to understand the alignment of this administrative records (AR) information with respondent collected data.

The MAF-ARF is an internal Census Bureau administrative records file containing all known addresses for each individual within Census products. Census gathers residence information from several sources, including but not limited to the Internal Revenue Service, the Department of Housing and Urban Development, and the U.S. Postal Service. This information allows the tracking of respondent location over time. Established in the 1970s, the NSCG is a biennial survey that collects data on the college-educated population of the United States, highlighting the connection between educational attainment and subsequent labor market outcomes.

This study evaluates conceptual alignment, linkage, and agreement of residential location information between the two data sources and their stability across certain respondent characteristics. Upon linking the MAF-ARF to the NSCG by Protected Identification Key (PIK), the data shows generous coverage of the NSCG sample, on average over 90 percent. The agreement rate for both sources reporting the same address varies over time. The maximum is 84.80 percent in 2011, and then declines over time to 59.38 percent in 2015. Between the 2010 and 2013 surveys, the MAF-ARF predicts just over half (54.54 percent) of non-returning respondents did not move from their initially reported address. This suggests a significant issue with respondent refusals to participate in future surveys. Among returning respondents the MAF-ARF and NSCG predict migration rates of 34.90 and 39.51 percent, respectively. Mobility appears to be most correlated with respondent age, declining from 73.47 percent among those in their twenties to 27.02 percent among those aged 65 and older.

One important limitation of this AR source is a methodological change in its production in 2012. At that time, the MAF-ARF discontinued reporting multiple addresses for a person and began reporting single addresses. This change was implemented for privacy protection, randomizing the selection of a unique address among several locations meeting a minimum threshold of confidence in accuracy. The reduction in available addresses may in turn reduce the probability of successfully re-establishing contact with respondents for follow-up interviews.

Nevertheless, the MAF-ARF is a trusted resource consistently used in research, and constitutes a good initial reference for locating an individual. Future research should investigate using other administrative records (including IRS 1040 and 1099 data), to construct residential history files for 2010 and 2013 NSCG respondents that can be compared to location information from the MAF-ARF and with the NSCG. A more detailed residential history file has the potential to improve NSCG operations while overcoming the limitations of the MAF-ARF.

Table of Contents

List of Tables

I. INTRODUCTION

The National Survey of College Graduates (NSCG) is a longitudinal survey of the college-educated population living in the United States. Sponsored by the National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation (NSF), the survey informs two congressionally-mandated reports, *Women, Minorities, and Persons with Disabilities in Science and Engineering* and *Science and Engineering Indicators*, on the composition and productivity of the nation's STEM workforce. Thus, NCSES, with the Census Bureau serving as the data collection contractor, administers the NSCG to collect information on the human capital investment decisions and labor market outcomes of highly-educated workers. Over time, the survey tracks respondents' demographic characteristics, educational attainment, workplace training, job satisfaction, professional mobility, and income.

NCSES is interested in the use of administrative data to improve the NSCG. Administrative records have the potential to address many goals, including: informing on measurement error, supplementing respondent-collected data, and reducing data collection and processing costs. To that end, NCSES has requested that the Demographic Research Area in the Center for Economic Studies at the Census Bureau (CES-Demo, formerly the Center for Administrative Records Research and Applications) evaluate the NSCG for the use of administrative records to enhance NSCG operations and data quality.

The NSCG employs a rotating panel design allowing researchers to conduct longitudinal analyses of human capital investments and labor market outcomes. Ideally, researchers using the NSCG would have access to education and employment information at four time points in six years. However, not all respondents complete all waves of data collection. One explanation for nonresponse is the inability of survey administrators to locate all respondents.

The Census Bureau's annual Master Address File Auxiliary Reference file (MAF-ARF) is a potentially useful resource to address survey nonresponse within NSCG. The Census Bureau populates the MAF-ARF with residence information from the United States Postal Service and other administrative records to track all known addresses for an individual with an assigned personal identifier. Each address is assigned a Master Address File Identification number (MAFID).

Linking the MAFID data from the MAF-ARF file to the NSCG master file, allows for the estimation of migration patterns for all baseline survey respondents. We will conduct separate analysis for both returning survey respondents and non-returning respondents. Our results will provide information that allows NSCG administrators to anticipate the need to send communications to alternative locations or via other modes of communication. This knowledge may potentially enhance survey operations by increasing the probability of successfully establishing contact with survey respondents, therefore improving efficiency and reducing data collection costs.

II. LITERATURE REVIEW

The purpose of this study is to evaluate the MAF-ARF as a tool to increase respondent retention within the NSCG. Between the 2010 and 2013 panels of the NSCG, 19.59 percent of the eligible

2010 sample did not participate in the 2013 follow-up survey.[1] Respondent attrition (a form of unit nonresponse) occurs for many reasons, one of which is the inability to locate a respondent. Attrition is a growing issue among panel datasets, and is becoming more common within the U.S. (Massey and Tourangeau, 2013; BLS; Tourangeau, 2004).

The statistical implications of attrition are non-negligible – attrition potentially increases bias in survey and model estimates, as the reduced sample does not accurately reflect the population at large (Watson and Wooden, 2006). For example, Zabel (1998) finds that estimates of wage elasticity in labor supply models could be biased when failing to account for nonrespondents' heightened sensitivity to changes in wage. He finds that attritors are less attached to the labor force, displaying larger variation in employment status, hours worked, and wages.

Many studies agree attrition is not a random event and is positively correlated with certain respondent characteristics. Men, the youth and elderly, racial and ethnic minorities, the unmarried, lesser educated, renters, and urban residents all tend to have lower response rates than their counterparts (Watson and Wooden, 2006; Behr, Bellgardt and Rendtel, 2005; Zabel, 1998). Item nonresponse in a previous interview or survey, particularly on income and sensitive items, is a statistically significant predictor of future unit nonresponse (Loosveldt, Pikery and Billiet, 2002). Additionally, should a respondent return to a survey after missing a wave, it is important to consider the findings of Tourangeau, Groves and Redline (2010) who present empirical evidence on the direct relationship between nonresponse bias and measurement error in survey responses. This result indicates inconsistency in the quality of responses of nonrespondents, possibly fluctuating in response to perceived importance of and overall experience with a survey.

Researchers have argued that the driving causes of attrition are respondent refusals and the inability to locate (Massey and Tourangeau, 2013; Laurie et al., 1999). Respondents will refuse to participate in a survey for a variety of reasons including temporary circumstances that make participation an inconvenience, perceived intrusion of privacy, and unpleasant experience with the interviewer or survey instrument. As for the inability to locate a respondent, survey administrators may face issues with bypassing security at residences, scheduling conflicts, and outdated contact information most likely the result of relocation (Massey and Tourangeau, 2013; Zabel, 1998). As a result, suggested solutions focus on methods for minimizing inconvenience, developing rapport with participants, and effectively tracking respondents over time.

For panel studies, accurate information on the residence of respondents over time is paramount. Relocation does not necessarily mean the respondent is not willing to continue participation in a survey (Lepkowski and Couper, 2002). Therefore, survey administrators should make efforts to adjust the frequency and timing of contact between surveys, use mailings with address forwarding services, and utilize administrative records where available all in order to maintain up-to date location information and minimize nonresponse (BLS, Massey and Tourangeau, 2013, Laurie, Smith and Scott, 1999). In the event contact cannot be established, the use of refreshment

[1] Since the NSCG employs a rotating panel, only respondents sampled from the 2009 ACS were eligible for participation in the following 2013 survey. The remaining 2010 respondents sampled from the 2003 NSCG, and 2001, 03, 06 and 08 Survey of Recent College Graduates (RCG) rotated out of the sample. The 2013 sample was replenished by sampling from the 2011 ACS and 2010 RCG.

samples, as done with the rotating panel design employed by NSCG, are a suitable solution for adjusting for bias due to attrition (Deng et al., 2013).

III. DATA

3.1 National Survey of College Graduates (NSCG):

The NSCG is a biennial survey sponsored by the National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation, administered by the Census Bureau, and sampled from the American Community Survey (ACS). It uses a rotating panel design in which respondents answer questions about their employment status, earnings, and education up to four times over a period of about six years. One of the unique features of the NSCG is its collection of data on more subjective information such as motivating factors for the individual's human capital investments, change in career or employment status. Additionally, the data collected in this survey informs two congressionally mandated reports on the U.S. STEM labor force: *Women, Minorities, and Persons with Disabilities in Science and Engineering*, and *Science and Engineering Indicators*. Survey respondents are college graduates, living in the U.S., up to age 75.

This study uses 2010 and 2013 NSCG restricted access data. The 2010 survey was the first data release after switching to its current sample frame, the ACS. To maintain the continuity of the rotating panel design, 46,828 new observations from the 2009 ACS were added to the sample already including 30,360 return respondents sampled from the 2001-2008 panels of the National Survey of Recent College Graduates and the 2003 NSCG for a total of 77,188 observations. Just under half of the 2010 sample (48.78 percent) returned for follow-up interviews in the 2013 NSCG.

3.2 Master Address File Auxiliary Reference File (MAF-ARF):

The Master Address File Auxiliary Reference File (MAF-ARF) is an internal Census Bureau resource containing all known addresses for each individual within Census products. Census gathers residence information from several sources, including but not limited to the Internal Revenue Service, the Department of Housing and Urban Development, and the U.S. Postal Service.[2] Each address receives a Master Address File Identifier (MAFID). Likewise, each person has an assigned Protected Identification Key (PIK), which is used to link an individual's information across surveys and administrative records. The PIK is based on personal identification information such as Social Security Number, birth date, and name. Census' Person Identification Validation System employs a probabilistic matching algorithm to attach PIKs and MAFIDs to person and address information as outlined in Wagner and Layne, 2014. The MAF-ARF is available annually from 2000 to 2018. This study uses the 2009-2013 releases of this dataset. Each record in the MAF-ARF file represents a unique PIK-MAFID combination.

[2] The sources of information used to identify individuals' residence are as follows: the Census Numident, the Census Unedited File, the IRS 1040 and 1099 files, the Medicare Enrollment Database (MEDB), Indian Health Service database (IHS), Selective Service System (SSS), and Public and Indian Housing (PIC) and Tenant Rental Assistance Certification System (TRACS) data from the Department of Housing and Urban Development, and National Change of Address data from the US Postal Service United States Postal Service (Finlay, 2016)

3.3 Limitations

There are some limitations to the usability of the MAF-ARF for locating survey respondents. As mentioned previously, the structure of the MAF-ARF file has changed over time. Prior to 2012, the MAF-ARF listed all known MAFIDs for each PIK, which is potentially useful when trying to track and assess residential agreement for particularly mobile populations. The availability of multiple MAFIDs temporarily addressed a second limitation associated with the timing of respondent moves relative to data collection schedules. Starting in 2012, the MAF-ARF switched to unique PIK-MAFID observations for disclosure avoidance purposes. An additional limitation is that some addresses such as P.O. boxes and locations outside of the U.S. do not receive MAFIDs. This is an issue of conceptual misalignment between the MAF-ARF and the NSCG in which the survey requests the best address by which to contact the respondent, which may not be the residential address captured by the MAF-ARF.

IV. RESEARCH QUESTIONS

The research questions are as follows:

1. To what extent is the location information collected by NSCG conceptually *aligned* with the administrative record information?
2. How often do NSCG records *link* to administrative record data that can be used to supplement survey information?
3. How often do data from the administrative records source *agree* with the survey data by major subpopulation characteristics?

V. METHODOLOGY

The research questions of the previous section correspond to three analytical objectives of this research. That is, to assess linkage, conceptual alignment, and agreement of residential address information between the NSCG and MAF-ARF. This section presents supplemental information on the analysis used to produce the data in the results section.

5.1 Conceptual Alignment:

For research question #1, evaluation of conceptual alignment involves verifying the data collected within both data sources are as similar as possible. In this study, that involves an assessment of the type of information available within both data sources and determination of how well the administrative data supports the needs of the survey. This discussion is available within the results section. Additionally, the data management section outlines steps taken to make the data comparable for appropriate assessment.

5.2 Linkage:

For research question #2, unique person identifiers (PIKs) are assigned to each respondent, allowing linkage between NSCG and MAF-ARF data. PIKs allow linkage of information for a particular person across various Census surveys and administrative records. PIKs assignment to datasets occurs via the Person Identification Validation System (PVS), a probabilistic matching algorithm used to anonymize incoming data at the Census Bureau. This process uses personally identifiable information (PII) from the survey such as name, age, and address to search reference files containing all known transactions for an SSN. Once matching information is found in the reference files with a certain threshold of confidence, the unique PIK value replaces PII found on the survey data file.[3] The linkage rate, based on unique PIKs, represents the proportion of the PIKed NSCG sample found in the MAF-ARF, is calculated for each year from 2009 to 2015 and by respondent follow-up status (returning and non-returning).

5.3 Agreement:

Next, for research question #3, the analysis includes findings on 1) agreement in response value, and 2) variation in measurement error across respondent characteristics. For this analysis, the 2010 NSCG sample consists of two groups: respondents who participated in both the 2010 and 2013 surveys (returning), and those who did not return for the 2013 survey (non-returning). In the results section, Table 2 presents the agreement rate between the initial NSCG MAFID based on reported street address information and the MAF-ARF MAFID for a PIK across several years. The underlying frequencies of agreement between initial NSCG MAFIDs and MAF-ARF MAFIDs across the years 2009-2013 are available in the appendix. Table 3 provides the respondent 2010-2013 migration rate, as determined by a change in MAFID in the MAF-ARF. The findings of both Tables 2 and 3 are presented by respondent follow-up status (returning and non-returning). Table 4, displays a cross-tabulation of the 2010-2013 migration rate among returning respondents, as determined by changes in MAFID within the MAF-ARF and changes in the street address information reported to NSCG between the 2010 and 2013 surveys. Finally, Table 5 assesses the migration rate across respondent demographics and education characteristics for returning and non-returning respondents. Note that the migration rate for non-returning respondents is based only on MAF-ARF information as street address data was not available for this group in the 2013 master file. Its results predict how frequently a non-returning respondent should be available at a known initial address.

VI. DATA MANAGEMENT

This section describes the steps taken to prepare the data for analysis via the appendage of PIKs and MAFIDs to the NSCG sample, their subsequent linkage to the MAF-ARF, and assessment of MAFID agreement relative to the initial address information provided in the 2010 NSCG master file. Using PII available within the master file, the dataset underwent PVS processing to append PIKs for linkage to the MAF-ARF and MAFIDs for evaluation of location agreement. This processing resulted in a 98.52 percent PIK rate and a 95.40 percent MAF-match rate. Survey data typically has high PIK rates (90-93%) and failure to receive a PIK often occurs when the SSN is unknown and/or disconnected from government programs and records (NORC, 2011). However, since individuals of high socioeconomic status, such as college graduates, exhibit fewer of these characteristics, the higher than average PIK rate for the 2010 NSCG is justified. The rate of MAFID assignment is lower due to the inability to assign identifiers to P.O. boxes and locations outside of the U.S. Ultimately, this research focuses on the 2010 NSCG respondents sampled from the 2009 ACS and consists of approximately 45,000 unique PIKs.

Next, the assigned PIKs were appended to the 2010 NSCG response file via the survey unique identifier, REFID. After identifying unique PIKs from the NSCG file, it was linked to the 2009 MAF-ARF, which is a person-address-level file. Recall, that the 2009-2011 files contain multiple

[3] See Wagner and Layne (2014) for a detailed description of the PVS process.

MAFIDs per PIK, resulting in a one-to-many match when linking to the 2010 NSCG. Linkage to the 2012-2015 MAF-ARF were one-to-one matches.

To assess agreement between the NSCG and MAF-ARF MAFIDs, a binary indicator was coded for respondent immobility captured by the MAF-ARF. It took on a value of one if the NSCG MAFID equaled the MAF-ARF MAFID designating the respondent as a MAF-ARF non-mover. If the MAFIDs did not agree, the respondent was designated a MAF-ARF mover. This process was repeated for each MAF-ARF year used in the analysis. This was an iterative process for the years 2009-2011 requiring collapse around unique PIK values after determining if any of the MAF-ARF MAFIDs for a PIK were a match.

Furthermore, a separate binary indicator was coded to reflect immobility within the NSCG information. The indicator variable took on a value of one if the street address information in both the 2010 and 2013 NSCG master files were identical, signaling *non*-moving status. Otherwise, the respondent was designated an NSCG mover. The NSCG migration indicator was then linked to the NSCG subsample by REFID. Updated address information was only available for returning respondents in the 2013 master file. The frequency of values of these MAF-ARF and NSCG migration indicators across respondent follow-up status and personal characteristics are presented in the following results section.

VII. RESULTS

7.1 Conceptual Alignment

The NSCG collects address information from respondents in order to establish future communication with them for follow-up interviews. Note the address is presumably the "best" location to reach the respondent, and not necessarily an actual residence. Therefore, the analysis begins with using the street address data in the 2010 NSCG master file to assign initial MAFIDs for each respondent. The MAF-ARF provides MAFIDs for residential locations associated with an individual. In earlier years, all known locations are available, and later on just a single location per person assigned with some threshold of confidence.

Both sources are conceptually aligned in the provision of valid location information for the respondent. However, there is a distinction between "best" and valid contact information. For example, some respondents provide P.O. Box addresses to NSCG, which would not be found in the MAF-ARF. Even if the respondent provides a residential address, it may be that of a parent/relative or some other location the respondent spends a lot of time. The MAF-ARF only shows a valid address identifier for a person, meaning administrative records report the individual conducted a transaction some time that year attached to that location. In either of these cases, MAF-ARF may not provide the respondent's preferred address.

7.2 Linkage

Table 1 displays the linkage rate by PIK across several years of data. These percentages highlight the coverage of the NSCG subsample by the MAF-ARF information. For example, 92.18 percent of the NSCG subsample had address information available within the 2009 MAF-ARF. The linkage rate increases over time up to 2012 before dropping to its lowest rate in 2013. The reasons for this abrupt decline are unclear, although it appears to be related to a structural break that may indicate changes in the underlying methodology of the MAF-ARF, which switched from multiple to unique MAFIDs in 2012.

The pattern of the results persists within the returning and non-returning subgroups. However, notice that non-returning respondents consistently link less frequently to the MAF-ARF than returning respondents. This raises the question of whether the same characteristics associated with survey nonresponse are similar to those of individuals less frequently picked up in administrative data.

Table 1: PIK Linkage Rate to the MAF-ARF by 2013 NSCG Respondent Status

	Year							Total
Status	2009	2010	2011	2012	2013	2014	2015	Obs.
Returning	92.22	92.41	92.74	93.34	88.31	92.38	92.64	36,500
Non-Returning	91.98	92.18	92.36	92.98	82.08	88.56	87.73	8,700
Overall	**92.18**	**92.36**	**92.67**	**93.27**	**87.10**	**91.64**	**91.69**	**45,000**

Source: 2010 NSCG and 2009-2015 MAF-ARF files.

Note: Frequency values rounded for disclosure avoidance.

7.3 Agreement

Table 2 shows the overall agreement rate between NSCG and MAF-ARF MAFIDs for the subsample. Specifically, the percentages in this table indicate each year how frequently the MAF-ARF suggests the respondent did *not* move away from the address he initially reported in the 2010 NSCG. For example, according to the MAF-ARF, in 2011 84.40 percent of the sample were still located at the same address as in the 2010 NSCG. The 2011 MAF-ARF yields the maximum agreement rate before declining over the remaining years. It is expected that respondent mobility increases over time, resulting in lower agreement rates. However, a clear driver of the observed decline in these results is the switch to unique MAFID assignment after the 2011 MAF-ARF. Not only does the rate of agreement decrease dramatically, but also the quality of the match does not recover over time. Lastly, as seen with the linkage results, non-returning respondents underperform relative to returning respondents.

Table 2: Rate of Agreement on Location by 2013 NSCG Respondent Status

	Year							Total Obs.
Status	2009	2010	2011	2012	2013	2014	2015	
Returning	73.07	80.07	85.30	71.35	65.10	64.32	61.38	36,500
Non-Returning	69.14	76.59	82.77	66.41	54.54	54.59	51.04	8,700
Overall	**72.31**	**79.40**	**84.80**	**70.39**	**63.06**	**62.44**	**59.38**	**45,000**

Source: 2010 NSCG and 2009-2015 MAF-ARF files.

Note: Frequency values rounded for disclosure avoidance. Agreement occurs when the MAFID associated with the respondent-reported address in the 2010 NSCG matches the MAFID reported for that individual in the MAF-ARF.

The underlying frequencies used to derive the overall agreement rates can be used to trace mobility status over time. Specifically, one can follow a PIK each year and observe whether or not the person moved from their initial NSCG address, and if so, whether they returned later.

This very well could be the case for young adults that rely on parents' homes as a safety net for temporary residence. The table in the appendix tracks MAF-ARF mobility status over time and, for each year, segments frequencies of the previous year MAF-ARF movers and nonmovers into current movers or nonmovers. This exercise shows that according to the MAF-ARF 44.85 percent of the sample never moved from the initial NSCG address over the years 2009-2015. Likewise, 11.29 percent of the sample did not live at the initial NSCG address at any time between 2009 and 2015.[4]

Within the context of biannual survey operations, the migration rate between waves of the survey is particularly relevant. Focusing on migration between 2010 and 2013, Table 3 shows the MAF-ARF predicts over half (54.54%) of non-returning respondents could be reached at their initial NSCG address.

Table 3: 2010-2013 Migration Rate by 2013 NSCG Respondent Status

		2013 MAF-ARF Status		
		Moved	Did Not Move	Total
Respondent Status	Did Not Return	4,000	4,800	8,700
		45.46	54.54	100.00
	Returned	12,500	23,500	36,500
		34.90	65.10	100.00
	Total	16,500	28,500	45,000

Source: 2010 and 2013 NSCG and 2009-2015 MAF-ARF files.

The following table allows comparison of MAF-ARF migration rates to those derived from the survey information directly. As mentioned in the data management section, MAFIDs were also assigned to the 2013 NSCG master file based on available street address information. Since this data was only available for returning respondents, Table 4 focuses on that subsample. The NSCG shows 39.51 percent of return respondents changed addresses between 2010 and 2013. Similarly, the MAF-ARF indicates 34.90 percent moved during that same time. Both data sources agree the respondent moved 28.78 percent of the time. Summing along the diagonal reveals the overall agreement in migration status (both mover and non-mover) between the NSCG and MAF-ARF is 83.15 percent.

[4] Divide the top right value (20,255) of the appendix table by the sample size (45,166) to obtain percentage of sample that never moved. Divide the bottom right value (5,097) of the table by the sample size (45,166) to obtain the percentage of the sample that never lived at the initial NSCG address.

Table 4: NSCG/MAF-ARF Agreement on Mobility 2010-2013 among Return Respondents

		MAF-ARF		
		Moved	Did Not Move	Total
NSCG	Moved	10,500	3,900	14,500
		28.78	10.73	**39.51**
	Did Not Move	2,200	20,000	22,000
		6.12	**54.37**	60.49
	Total	12,500	23,500	36,500
		34.90	65.10	100.00

Source: 2010 and 2013 NSCG and MAF-ARF files.

Table 5 explores variation in the migration rate across respondent characteristics. The first three columns are results for returning respondents, while the last two columns display results for non-returning respondents. Among returning respondents, the MAF-ARF estimates are consistently conservative relative to the NSCG. The most noticeable variation in the migration rate occurs across age groups. Specifically, mobility decreases with age; a reasonable result. The migration rate for respondents in their twenties is almost double the overall rate—73.47 vs. 39.51 percent for NSCG and 67.24 vs. 34.90 percent for MAF-ARF. There appears to be greater migration among smaller racial groups such as Pacific Islanders and Native Americans. Women also seem to move more frequently than men do. Towards the end of the table, migration results for respondents with formal education in STEM disciplines do not appear to differ from those without STEM education. Among non-returning respondents, the patterns remain consistent. However, the MAF-ARF estimates are higher for this group than the analogous results for returning respondents, highlighting the possibility that the non-returning group have less stable living situations.

Table 5: Migration Rate across Respondent Characteristics

	Returning Respondents			Non-Returning Respondents	
	Count	NSCG	MAF-ARF	Count	MAF-ARF
Overall	**36,500**	**39.51**	**34.90**	**8,700**	**45.46**
Gender					
Male	20,000	38.92	33.90	4,700	45.07
Female	16,500	40.25	36.14	4,000	45.91
Age					
21-29	3,900	73.47	67.24	1,200	68.37
30-39	8,300	52.58	43.61	2,300	54.33
40-49	8,500	32.36	28.14	1,900	39.78
50-64	12,000	28.39	26.72	2,100	36.91
65-75	3,600	27.02	23.42	1,200	30.37
Race					
White	24,000	37.79	32.61	5,100	42.94
Black	3,000	39.47	37.63	1,000	45.11
Asian	5,400	43.92	38.56	1,500	51.46
Pacific Islander	150	43.75	56.25	50	42.00
Native American	150	45.81	54.19	50	57.45
Multiple race	3,700	43.80	41.20	1,000	49.07
STEM Education					
STEM	20,500	39.58	34.48	4,900	46.14
Non-STEM	15,500	39.42	35.47	3,900	44.59

Source: 2010 and 2013 NSCG and MAF-ARF files.

VIII. CONCLUSION

Summary of Results

This research evaluates conceptual alignment, coverage, and agreement of residential address information between the National Survey of College Graduates (NSCG) and Master Address File-Auxiliary Reference File (MAF-ARF) data sources. These datasets were linked by PIK in order to achieve a person-address-level file from which to compare location information from several administrative data sources. The MAF-ARF provides good coverage of the NSCG sample over time, as high as 93.27 percent in 2012. The average linkage rate over the years 2009-2015 is 91.56 percent.

Looking at the agreement rate between the linked datasets over time uncovered structural limitations of the MAF-ARF. The agreement rate on respondents not changing addresses rises from 2009 to 2011 as high as 84.80 percent. That rate abruptly declines to 70.39 percent in 2012 and trends downward through 2015. The initial decline in the agreement rate corresponds to the year in which the MAF-ARF discontinued reporting multiple possible addresses for a person.

After 2011, only one address per person was provided, greatly reducing the probability of agreement with the original address reported to NSCG in 2010. Overall, the MAF-ARF predicts 44.85 percent of the sample never moved from their initially reported address from 2009-2015. Likewise, 11.29 percent did not there from 2009-2015.

The following results focus on migration between 2010 and 2013, assessing the availability of 2010 respondents for the 2013 survey. Most surprisingly, the MAF-ARF predicts that over half (54.54 percent) of non-returning respondents were still living at their initially-reported address. This result aligns with documented conclusions in unit nonresponse literature that the greater issue with attrition is with respondent refusals to continue participation as opposed to inability to locate them. Among respondents that continued their participation into the 2013 survey, the MAF-ARF and the NSCG agree on the respondents mobility status 83.15 percent of the time. Also, the predicted migration rates between the datasets are similar. The MAF-ARF indicates 34.90 percent of returning respondents changed addresses between 2010 and 2013, compared to the NSCG reporting 39.51 percent. The MAF-ARF typically reports more conservative estimates of the migration rate relative to the NSCG, even across respondent characteristics. Also, the MAF-ARF indicates higher migration rates for non-returning respondents than for returning respondents. The greatest variation in migration occurs across age groups where the youngest are most mobile (73.47 percent) and the oldest the least (27.02 percent).

Recommendations for Future Work

The methodological change in the production of the MAF-ARF presents pros and cons for its usage. In earlier years, the abundance of information with multiple addresses provided several locations surveys could use to successfully locate respondents or communicate with someone that knows an accurate address. The unique addresses provided in more recent years meet certain thresholds of confidence for accuracy. However, it is not well understood (intentionally so, for disclosure avoidance) how the reported address is chosen over other threshold surpassing addresses linked to an individual at a point in time. Ideally, a survey would have as much information as possible available to carry out its operations. Nevertheless, the MAF-ARF is a trusted resource consistently used in research and constitutes a good initial reference for locating an individual.

The alternative administrative source of location information would be IRS 1040 and 1099 forms, which are an input into the production of the MAF-ARF. With this data, it may be possible to evaluate the accuracy of MAF-ARF results based on proximity to work and the state in which this information is being filed. It can also provide contact information on others who may know where to locate a respondent, such as parents/legal guardians, a spouse, or business partner.

Additional administrative records data, including IRS 1040s and information returns, may provide a better measure of mobility among this sample of college graduates. To investigate whether this may yield improvements when combined with the existing MAF-ARF data, we propose the construction of a residential history file for 2013 and 2015 respondents combining information from the MAF-ARF and IRS 1040 and 1099 records. With this dataset, we can derive summary statistics on migration rates and conduct analyses on the determinants of migration.

IX. REFERENCES

Behr, A., E. Bellgardt and U. Rendtel. 2005. "Extent and Determinants of Panel Attrition in the European Community Household Panel". *European Sociological Review*. Vol. 21 no. 5. 489-512.

Deng, Yiting, D. Sunshine Hillygus, Jerome P. Peiter, Yajuan Si and Siyu Zheng. 2013. "Handling Attrition in Longitudinal Studies: the Case for Refreshment Samples". *Statistical Science*. Vol. 28 no. 2. 238-256.

National Longitudinal Surveys of the U.S. Bureau of labor Statistics. National Longitudinal Survey of Youth 1997: Retention & Reasons for Non-Interview. Accessed Oct. 18, 2018. https://www.nlsinfo.org/content/cohorts/nlsy97/intro-to-the-sample/retention-reasons-non-interview/page/0/1

Laurie, H., R. Smith and L. Scott. 1999. "Strategies for Reducing Nonresponse in a Longitudinal Panel Survey". *Journal of Official Statistics*. Vol. 15 no 2. 269-282.

Lepkowski, J. M. and M. P. Couper. 2002. "Nonresponse in the Second Wave of Longitudinal Household Surveys". In *Survey Nonresponse*, eds. R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little. John Wiley & Sons, New York.

Loosveldt, Geert, Jan Pickery and Jaak Billiet. 2002. "Item Nonresponse as a Predictor of Unit Nonresponse in a Panel Survey". *Journal of Official Statistics*. Vol. 18 no. 4. 545-557.

Massey, Douglas S. and Roger Tourangeau. 2013. "Where Do We Go from Here? Nonresponse and Social Measurement". *The ANNALS of the American Academy of Political and Social Science*. Vol. 645 no. 1. 222-236.

Tourangea, Roger. 2004. "Survey Research and Societal Change". *Annual Review of Psychology.* Vol. 55. 775-801.

Tourangeau, Roger, Robert M. Groves and Cleo D. Redline. 2010. "Sensitive Topics and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error". *Public Opinion Quarterly*. Vol. 74 no. 3. 413-432.

Wagner, Deborah and Mary Layne. 2014. "The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software". CARRA Working Paper no. 2014-01. U.S. Census Bureau, Washington, D.C.

Watson, Nicole and Mark Wooden. 2009. "Identifying Factors Affecting Longitudinal Survey Response". In *Methodology of Longitudinal Surveys*, ed. Peter Lynn. John Wiley & Sons, New York.

Zabel, Jeffrey E. 1998. "An Analysis of Attrition in the Panel Study of Income Dynamics and the Survey of Income and Program Participation with an Application to a Model of Labor Market Behavior". *The Journal of Human Resources*. Vol. 33 no. 2. 479-506.

X. APPENDIX

Table 6: Frequency of Change in Residence among Linked Cases

Note: white cells are counts of obs. where the 2010 NSCG MAFID matches the MAF-ARF MAFID (non-mover); shaded cells are counts where the 2010 NSCG MAFID does not match the MAF-ARF MAFID (mover); (D) used for substantially small counts for disclosure avoidance.

universe	2009	2010	2011	2012	2013	2014	2015
45,166	32,661	32,258	31,453	26,906	22,935	21,543	20,255
							1,288
						1,392	470
							922
					3,971	1,239	796
							443
						2,732	227
							2,505
				4,547	1,568	1,079	822
							257
						489	199
							290
					2,979	270	140
							130
						2,709	119
							2,590
			805	242	122	102	92
							10
						20	(D)
							(D)
					120	50	38
							12
						70	(D)
							(D)
				563	32	21	(D)
							(D)
						11	(D)
							(D)
					531	136	117
							19
						395	17
							378

Source: 2010 NSCG and 2009-2015 MAF-ARF

universe	2009	2010	2011	2012	2013	2014	2015
45,166 cont.	32,661 cont.	403	179	121	61	(D)	(D)
							(D)
						(D)	(D)
							(D)
					60	21	11
							10
						39	(D)
							(D)
				58	10	(D)	(D)
							(D)
						(D)	(D)
							(D)
					48	(D)	(D)
							(D)
						(D)	(D)
							(D)
			224	45	11	(D)	(D)
							(D)
						(D)	(D)
							(D)
					34	14	(D)
							(D)
						20	(D)
							(D)
				179	10	(D)	(D)
							(D)
						(D)	(D)
							(D)
					169	(D)	(D)
							(D)
						(D)	(D)
							(D)

Source: 2010 NSCG and 2009-2015 MAF-ARF

universe	2009	2010	2011	2012	2013	2014	2015
45,166 cont.	12,505	3,603	3,339	2,151	1,524	1,326	1,210
							116
						198	67
							131
					627	114	62
							52
						513	27
							486
				1,188	321	223	170
							53
						98	26
							72
					867	47	17
							30
						820	23
							797
			264	50	14	(D)	(D)
							(D)
						(D)	(D)
							(D)
					36	11	(D)
							(D)
						25	(D)
							(D)
				214	11	(D)	(D)
							(D)
						(D)	(D)
							(D)
					203	18	(D)
							(D)
						185	(D)
							(D)

Source: 2010 NSCG and 2009-2015 MAF-ARF

universe	2009	2010	2011	2012	2013	2014	2015
45,166 cont.	▨ 12,505 cont.	▨ 8,902	3,333	2,022	1,407	1,241	1,113
							▨ 128
						▨ 166	51
							▨ 115
					▨ 615	104	32
							▨ 72
						▨ 511	23
							▨ 488
				▨ 1,311	309	229	180
							▨ 49
						▨ 80	26
							▨ 54
					▨ 1,002	59	20
							▨ 39
						▨ 943	21
							▨ 922
			▨ 5,569	257	75	60	49
							▨ 11
						▨ 15	(D)
							▨ (D)
					▨ 182	42	27
							▨ 15
						▨ 140	10
							▨ 130
				▨ 5,312	70	44	(D)
							▨ (D)
						▨ 26	(D)
							▨ (D)
					▨ 5,242	98	69
							▨ 29
						▨ 5,144	47
							▨ 5,097

Source: 2010 NSCG and 2009-2015 MAF-ARF